REPORT DOCUMENTATION PAGE

Form Approved OMB NO. 0704-0188

The public reporting burden for this collection of information is estimated to average 1 hour per response, including the time for reviewing instructions, searching existing data sources, gathering and maintaining the data needed, and completing and reviewing the collection of information. Send comments regarding this burden estimate or any other aspect of this collection of information, including suggesstions for reducing this burden, to Washington Headquarters Services, Directorate for Information Operations and Reports, 1215 Jefferson Davis Highway, Suite 1204, Arlington VA, 22202-4302. Respondents should be aware that notwithstanding any other provision of law, no person shall be subject to any oenalty for failing to comply with a collection of information if it does not display a currently valid OMB control number.
PLEASE DO NOT RETURN YOUR FORM TO THE ABOVE ADDRESS.

1. REPORT DATE (DD-MM-YYYY)	2. REPORT TYPE Technical Report	3. DATES COVERED (From - To) -

4. TITLE AND SUBTITLE
Progress Report - March 2012

5a. CONTRACT NUMBER
W911NF-12-1-0037

5b. GRANT NUMBER

5c. PROGRAM ELEMENT NUMBER

6. AUTHORS
Alessandro Flammini

5d. PROJECT NUMBER

5e. TASK NUMBER

5f. WORK UNIT NUMBER

7. PERFORMING ORGANIZATION NAMES AND ADDRESSES
Indiana University at Bloomington
Trustees of Indiana University
509 E 3RD ST
Bloomington, IN 47401 -3654

8. PERFORMING ORGANIZATION REPORT NUMBER

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)
U.S. Army Research Office
P.O. Box 12211
Research Triangle Park, NC 27709-2211

10. SPONSOR/MONITOR'S ACRONYM(S)
ARO

11. SPONSOR/MONITOR'S REPORT NUMBER(S)
61766-NS-DRP.1

12. DISTRIBUTION AVAILABILITY STATEMENT
Approved for public release; distribution is unlimited.

13. SUPPLEMENTARY NOTES
The views, opinions and/or findings contained in this report are those of the author(s) and should not contrued as an official Department of the Army position, policy or decision, unless so designated by other documentation.

14. ABSTRACT
Progress by IU, UM, and ATL for SMISC DARPA project for march 2012

15. SUBJECT TERMS
geolocation, time series analysis

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	15. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON Alessandro Flammini
a. REPORT	b. ABSTRACT	c. THIS PAGE			19b. TELEPHONE NUMBER 812-856-1830
UU	UU	UU	UU		

Report Title

Progress Report - March 2012

ABSTRACT

Progress by IU, UM, and ATL for SMISC DARPA project for march 2012

DARPA SMISC Project:
DESPIC: Detecting Early Signatures of Persuasion in Information Cascades

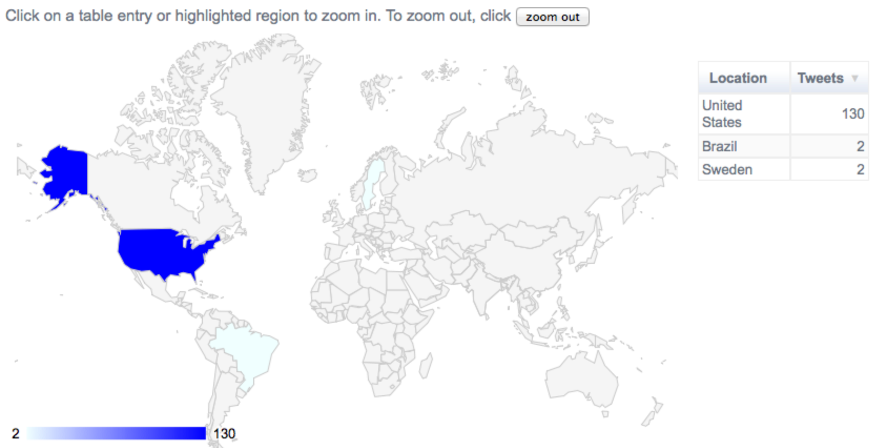
Teams:
Indiana University: A. Flammini (PI) and F. Menczer
University of Michigan: Qiaozhu Mei
Lockheed Martin Advanced Technology Laboratories (ATL): S. Malinchik

Progress Report – March 2012

General: The teams had their internal kick-off meeting in Bloomington, IN on March 1st-2nd. The discussion focused on three main issues:

- *High-level design of the data collection/detection infrastructure*. The goal is to have an infrastructure with components that can be developed as much as possible independently, has efficient underlying data storage capabilities (especially for time series) and allows a smooth the input/output exchange between the different components.
- *Possible approaches to the meme identification problem*. Since the UM team, originally charged with this task, will have not the possibility to hire a student until at least the summer semester, some initial exploratory work on this issue will be performed at IU, mostly through the work of the postdoc that will join the group in April.
- *Exchange of preliminary data in order for teams to start developing relevant infrastructures/code*. In particular IU will provide ATL a small number of time series of meme network features. As soon as resources will be available, UM will provide IU with a small number of tweets of memes relative to cases of rumor spreading.

IU: The IU team is working at expanding the visualization capabilities of the Truthy platform that will constitute the basis for the proposed infrastructure. Such visualization tools will help in the initial exploratory phase of suspicious meme detection, and in a second moment, the public use of data produced by our infrastructure. In particular we have been developing a sub-system that extracts geo-location relevant information (latitude/longitude) from tweets, determines the country of origin through a publicly available *Yahoo! API*, aggregates tweets over a period of time of 60 days, and shows the resulting information on a world map (see figure below).
Since the fraction of tweets carrying explicit geo-location information is very small (0.5% is our current estimate) we also considering introducing less accurate but larger-coverage heuristic methods to determine place of origin of tweets.


Click on a table entry or highlighted region to zoom in. To zoom out, click zoom out

Location	Tweets ▾
United States	130
Brazil	2
Sweden	2

ATL: ATL team has been focusing on exploring technologies to compare and cluster time series that represent multiple features of information cascades in social media. Below is a summary of our research.

The goal of time series clustering is to organize data into homogeneous groups where the within-group-object similarity is minimized and the between-group-object dissimilarity is maximized. Clustering is necessary when labeled data are not available.

One key component in time series clustering is the function used to measure the similarity between two time series being compared.

The vast majority of publications are focused on indexing time series under the Euclidean distance metric. However, there is an increasing awareness that the Euclidean distance is a very brittle distance measure. There is understanding that another distance measure is necessary allowing an elastic shifting of the time axis, to accommodate sequences that are similar but out of phase. *Berndt and Clifford (1994)* introduced the technique called *Dynamic Time Warping* (DTW). DTW is an algorithm for measuring similarity between two sequences that may vary in time or speed. The superiority of DTW over Euclidean distance has been demonstrated by many authors and there is increasing evidence that the classic Dynamic Time Warping measure is the best measure in most domains [2].

A challenging task of time series analysis is to find the best match to a query time series, from typically very large pool of candidates. This can trivially be achieved by sequential scanning, comparing each and every candidate to the query, in an arbitrary order. The problem with sequential scanning is that it is simply too slow for most applications. What we really need is a technique to index the data, that is, to find the best match without having to examine every candidate.

A classic trick to speed up sequential search with an expensive distance measure such as DTW is to use a cheap-to-compute lower bound (LB) to prune off unpromising candidates [2,3]. The strategy is to use a lower bounding calculation as often as possible and only do the expensive, full calculations when it is absolutely necessary. Keogh demonstrated that although DTW is $O(n^2)$ complexity, after setting the warping window for maximum accuracy, we only have to do 6% of all calculations, and if we use the LB_Keogh lower bound, we only have to do 0.3% of the work! [4]. Some rules to speed up calculations [4] include *Optimized Z-normalization* and use of multiple LB in a cascade for efficient pruning.

We are currently exploring two kinds of similarity: similarity at the level of shape and similarity at structural level. The technique described above is the best for the shape-based similarity. For long time series, shape based similarity will give very poor results. We need to measure similarity based on high-level structure. The basic idea is to extract *global* features from the time series (*Zero Crossing, Autocorrelation, ARIMA*, etc), create a feature vector, and use these feature vectors to measure similarity and/or classify. *Compression Based Similarity* is another option [5].

Software identified:
1. *SAX: Symbolic Aggregate approximation.* http://www.cs.ucr.edu/~eamonn/SAX.htm
2. *JMotif*: Implements algorithms for time-series analysis and data mining in Java and R. http://code.google.com/p/jmotif/

UM:
1. A graduate assistant is identified to work on this project in the summer (renewable in the Fall upon the performance).
2. Tweets about 5 selected rumors are extracted, and will be shared with the IU team shortly.
3. We are currently investigating the issue of retrieving rumor related tweets without supervision.

References:

[1] Berndt D, Clifford J, 1994. *Using dynamic time warping to find patterns in time series.* AAAI-94 workshop on knowledge discovery in databases, pp 229–248.
[2] H. Ding, G. Trajcevski, P. Scheuermann, X. Wang, and E. J. Keogh. 2008. *Querying and Mining of Time Series Data: Experimental Comparison of Representations and Distance Measures.* PVLDB 1, 2, 1542-52.
[3] E. Keogh, L. Wei, X. Xi, M. Vlachos, S.H. Lee, and P. Protopapas. 2009. *Supporting exact indexing of arbitrarily rotated shapes and periodic time series under Euclidean and warping distance measures*. VLDB J. 18, 3, 611-630.
[4] Rakthanmanon T, Campana B, Mueen A, Batista G, Westover B, Zhu Q, Zakaria J, Keogh E, 2012. *Searching and Mining Trillions of Time Series Subsequences under Dynamic Time Warping* (in press).
[5] E.Keogh, S.Lonardi, C.A.Ratanamahatana. *Towards Parameter-Free Data Mining*, Proceedings of *ACM Conference on Knowledge Discovery and Data Mining* (KDD'04), pp.206-215, Seattle, WA, 2004.